Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 70 (2015) 41 – 47

Procedia Computer Science

4th International Conference on Eco-friendly Computing and Communication Systems

Regression and Endogeneity Bias in Big Marketing Data

PankajDeep Kaur*, Sumedha Arora[a*]

Guru Nanakdev Univresity ,RC jalandhar,144001,India
Guru Nanakdev Univresity ,RC jalandhar,144001,India

Abstract

Big marketing data offers more interesting and challenging problems but along with greater opportunities. These days calibration is performed in the marketing field for supporting the managers in marketing-mix decisions. It is also done to create general knowledge that paves a way for better understanding of marketing relationships. Hence it indirectly supports decisions. The marketing data is adulterated with endogeneity and the regressions, both requiring optimizable response model. The models should always be implementable if actual decision support is the objective. Endogeneity can be removed with the help of structural equations. Owing to this endogeneity challenge it is difficult to understand how the managers can reach their decisions. Endogeneity removal allows improvement in managerial decision-making

Peer-review under responsibility of the Organizing Committee of ICECCS 2015

Keywords: regression ;endogeneity;decision;variables;

1. Introduction

In modern times, data-mining is a technique which gathers useful information about the customer and discovers the hidden customer's hidden behavior from big data. It also provides a great help by guiding decision- making and by forecasting the impacts of the decisions . Many marketing studies related with big marketing data focus mostly on determining the true effects of marketing instruments, such as personal selling or advertising. Such knowledge is needed to evaluate policies, assess performance, and optimize the marketing-mix. If the findings are cited properly, practitioners and academicians can learn from them and make prior preparation for assessments. Thus they can

Peer-review under responsibility of the Organizing Committee of ICECCS 2015
doi:10.1016/j.procs.2015.10.025

raise their standards. More and more fact-based studies are needed in Marketing science that can provide descriptive material on how managers behave. These data would create a challenge for developing theories of managers' behavior. Several contributions address this problem and examine factors leading to increased prices of certain commodities. It has been found that prices rise due to different factors. Bergstrom has discussed the disparities in pricing strategies employed by commercial builders. The precision-making framework aims at helping managers in identifying the potential characteristics of different customers belonging to divergent categories . Decision-making framework is known for providing the best precision-marketing strategy for companies. Owing to economic globalization and increase of competition in markets, economic pressures on the managers have forced them to face the problem of determining the policies of right strategic decision-making for selling the right products to the right customers at the right time.

The paper starts with the literature survey, throws light on endogeneity which is described in section 3, followed by regression handling for endogeneity removal by providing suitable model for it. Finally the experiment result analysis of the model is performed in section 5th.

2. Literature Survey

We can take the case when the sales people have to allocate calls across many of the small sales coverage units. This is particularly important when the data is to be reported for short time intervals ie in months. It is the fact that the optimal call frequency in sales is usually specified on the basis of a year. In such cases , the sales people usually decide which calls are convenient in particular months .They try to achieve optimal performance levels in the course of the whole year. Thus, it can be observed that there is a substantial randomness of calls across time but not across customers. As a result, critical levels of endogeneity for marketing spending data may be given only at more aggregate levels but not at disaggregate levels. This can take us on to the path of difficulties in the decision making , which is to be tackled as soon as possible. In order to deal with these problems, many decision-making techniques have been proposed in literature. Chen & Wang (2009) presented multi-criteria optimization and compromise solutions [1]. Saen (2010) developed a technique for order performance via similarity to ideal solution [2]. Hsu et al. (2010) developed a non-linear programming for decision-making, and Lin et al. (2011) presented a Linear programming for decision making [3]**.** Even customer's useful information and their hidden behaviours can be very easily extracted from Big Data. This ultimately helps in decision making and in forecasting the effects of decisions. More details of criteria used for decision- making by the managers have been depicted and discussed in following flow chart(Fig 1) and that ensuing paragraph.

2.1 Decision-making Framework

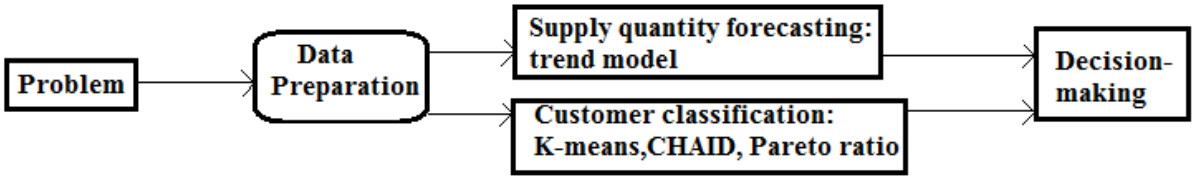

Fig. 1. (a) Decision making framework

(1) Data preparation: It is the process of preparing and analysing the quality of customers to determine their objectives (ref Fig1) (2) Supply quantity forecasting: It is used to predict the monthly supply quantity. It is done by using the appropriate model (3) Customer classification: It is the process of analysing every dimension of the customers' information by using a RFM model, decision trees algorithm, clustering algorithm and it also makes use of Pareto ratio to classify the customers' objectives; and (4) Marketing decision-making: It is used to make appropriate supply decisions based on the different categories of customer.

3. Endogeneity- The Concept

The endogeneity bias states that a variable is endogenous only if it is correlated with some error term. In principle, endogeneity is born in some static models due to the results of omitted variables, measurement errors, auto correlated errors. For example, endogeneity shows its impact in various economic applications, such as wages earned, hours worked, sale prices of commodities [4], etc. The endogeneity is tackled usually by adopting a control variable approach. The basic idea is to add a variable to the regression in such a way that, once a condition on this variable is applied, the regressors and unobservable go independent.

3.2 *Approaches For endogeneity*

3.2.1 Instrumental variable approach

One way to correct the endogeneity bias is to replace an endogenous independent variable with its estimates based on exogenous variables. Considering the following relation eq(1):

$$y = \alpha_0 + \alpha_1 * x_1 + \alpha_2 * z_1 + € \quad (1)$$

and x_1 is endogenous, so there is a need for exogenous variables that can correlate with the endogenous variable but they should not correlate with the error term. Denote this additional exogenous variable as (z_2). Thus, x1 has to be instrumented on the basis of all exogenous variables, the variables that are exogenous but correlated with y (z_1,) and the instrumental variable (z_1), which is correlated with x_1 but not with the error term. This is performed with the estimation shown below:

$$x_1 = b_0 + b_1 * z_1 + b_2 * z_2 + v_2 \quad (2)$$

now x_1 in Eq. (1) is then replaced with its predicted values estimated from eq. (2). The estimation can be performed with Two-Stage Least Squares (2SLS). The estimates are identical for linear functions that take v into Eq. (2) as an additional variable according to the control function approach. different types of instruments that can be used are-

- *Exogenous instruments.*
  These are those instrumental variables that are originally assumed to be exogenous but actually are not exogenous.
- *Lagged instruments*
  Sometimes researchers use lagged variables as instruments Its usage is appropriate as long as the data has no dynamic effects on prices
- *Latent instruments*
  If instrumental variables available are unobserved, it reflects the usage of latent discrete instrumental variables [5]

3.2.2Structural model approach

Another method for dealing with the endogeneity is to specify a structural model through which the endogenous variables can be explained by other variables, and a system of equations is estimated simultaneously [6].
If applied, the values for the sales and the elasticities are updated. As a result, solution converges to the fixed point of the optimal solution(cleared from eq 3,4).

$$optimal\ buget_{kint} = \frac{optimal\ allocation\ weight_{kint}}{\sum countries \sum products \sum activities optimal\ allocation\ weight_{kint}} * total\ budget \quad (3)$$

$$opt\ alloc = \frac{profit\ cont_{kint} * opt\ unit\ sales_{kint}(opt\ mkgt\ elast_{kint} + opt\ growth\ elast_{kint})}{\sum countries \sum products \sum activities optimal\ allocation\ weight_{kint}} * total\ budget$$

(4)

where opt=optimal, cont=contribution, elast=elasticity

4. Endogeneity In Marketing

Endogeneity in marketing data produces the wrong or incorrect results that can affect the outcomes of marketing decisions. The increasing number of reports stress the danger of the endogeneity bias

4.1 Handling Endogeneity Using Regression

Regression is the process of learning relationships between inputs and the respective outputs from example data, which enables predictions for novel inputs. The aim of regression is to find out the parameters of the model that minimise some error on the training examples. The generic scheme for parametric regression is depicted in Figure 2. The input to the regression algorithm is the training data and a set of algorithmic meta-parameters, including for instance learning rates. Each regression algorithm assumes a certain type of model, e.g. linear least squares assumes a linear model. The output of the algorithm is a vector of model parameters, which are determined by minimising an error measure on the training data.

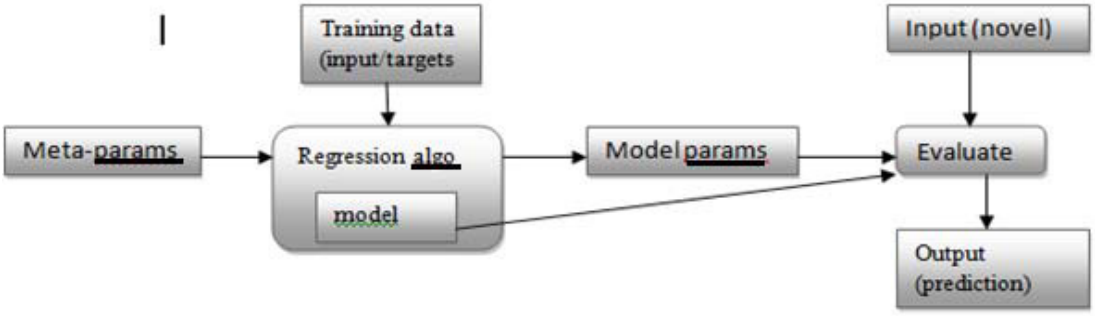

Fig. 2. Generic flowchart for parametric regression where algo=algorithm, params=parametric

4.2.Regression analysis in all over the market

The described factors led to the following regression [7] model:

$$price = \beta_0 + \beta_1 \text{factor}_1 + \beta_2 \text{factor}_1 + \beta_3 factor_1 + \sum \beta_{3+i} factor_1 + e \quad (5)$$

In the model eq (5) β0 is a constant while the β of the independent variables represent the magnitude in which the dependent variable price is influenced by the given independent variables. The error is represented by e.

5. Experiment And Result Analysis

This section integrate factors of house prices that are analyzed in a unified regression model based on current data. Tools used for performing this experiment can be weka. Discussing about the selling price of house. It can be considered that certain independent variables like age of house, the country size many other factors play important role in contributing a perfect model for dependent variable(SP). So in section 4.2 , the eq 5th can be reframed eq6.

$$House\ Price = \beta_0 + \beta_1 HouseAge + \beta_1 EnvRate + \beta_2 HouseSize + \beta_3 CountySize + \sum \beta_{3+i} Profit_1 + e \quad (6)$$

Table 1. House values for regression model

HouseAge	market	Size	Country size	profit	Selling price
3529	9191	6	0	0	$205,000
3247	10061	5	1	1	$244,900
4032	10150	5	0	1	$197,900
2200	9600	4	0	1	$195,000
3198	9669	5	1	1	$??/??

5.1 *Hypothesis about Factors of Influence for House Prices*

All the independent variables that are used in the models are as followed(considering the table 1)-

- *Size of house*-It is expected that a larger size of a housel leads to higher prices due to higher costs.
- *Country Size* - It is modelled by different variables (country).
- *Environmental rates*- Price differences among disciplines occur when looking at absolute price levels.
- *Profit status of builder*-It is an original characteristic of for-profit builder that they have a higher pressure to maximize financial results.
- *Age Of the house till now*- Age of house calculates the duration when the house was built with respect to the present date.

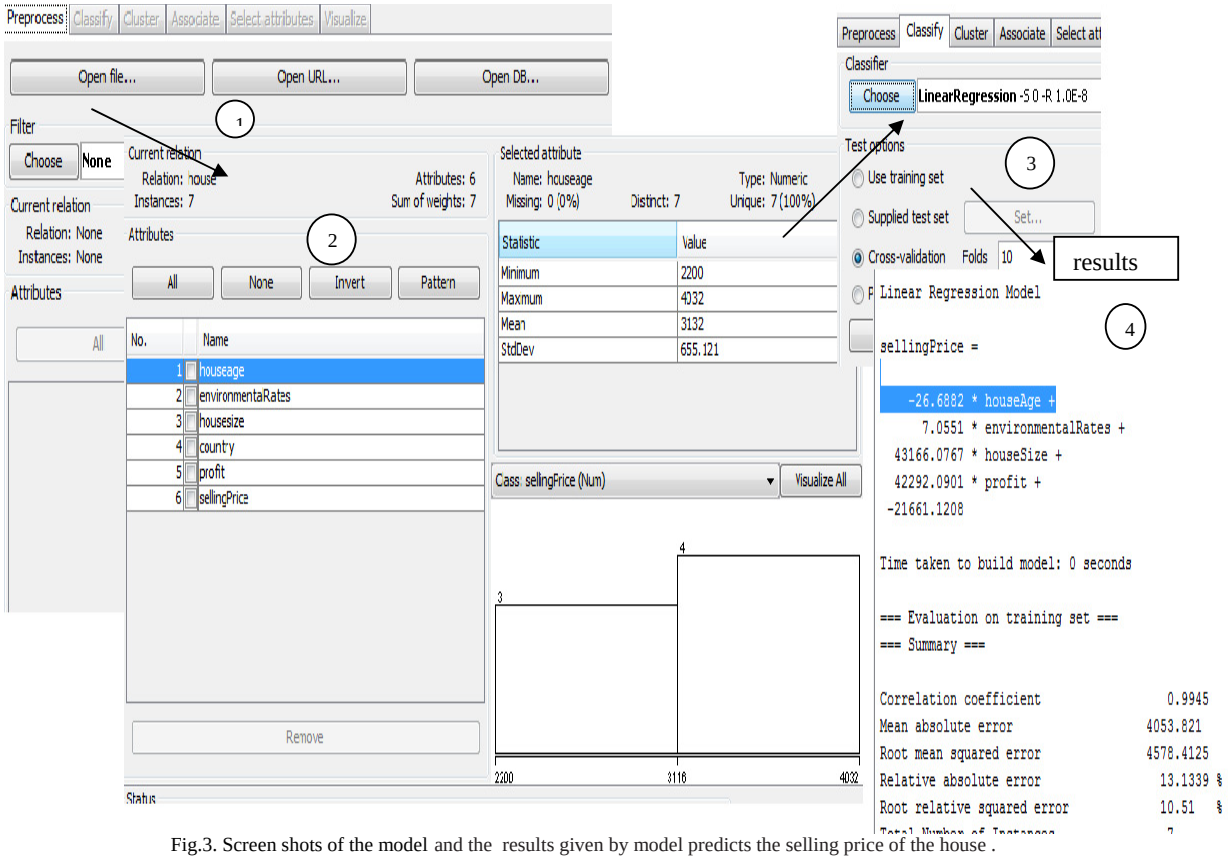

Fig.3. Screen shots of the model and the results given by model predicts the selling price of the house .

5.2. *Regression Analysis using Model*

The described factors led to the following regression model [8,9] .Referring Fig 3. By loading data into weka, regression model is created by using open file button. Once the file is choosed then right column shows that the maximum value of the data set for example(HouseAge) column is 4,032 square feet, and the minimum is 2,200 square feet average size is 3,131 square feet, standard deviation is 655 square feet, Finally Linear Regression leaf is selected for finding the linear regression. Supplied test set, Cross-validation helps to build a model.

Following are the **results** (encircled as 4) in Fig 3 and, shown above

- **Environmental rates**— WEKA only make use those columns that contribute towards the accuracy of the model. There by it ignores all those columns that are of no use in creating a good model.
- **profit do matter**—The model tells us it adds $42,292 to the house value.
- **Age of house houses reduce the value**— WEKA depicts that the older the house will be ,the lower will be the selling price of it. This is proved by the presence of the negative $26 coefficient in front of the house Age variable.
- **House Size do matter**— As The model tells us it adds $43,166 to the house value. There by with the increase in the size of the house, chances of house price will also increases.

This helps managers in decision making. After estimating a selling price from the above model, he can very easily categorise the customers by using Prato ratio in eq 7 As shown in the figure is how the customers can be categorised. For example considering the third category (Fig 4), with 9449 customers, accounted for 39% of all customers and its Pareto ratio is 0.9743, where Prato ratio is

$$pareto\ ratio = \frac{the\ percentage\ of\ marketing}{percentage\ customers} \tag{7}$$

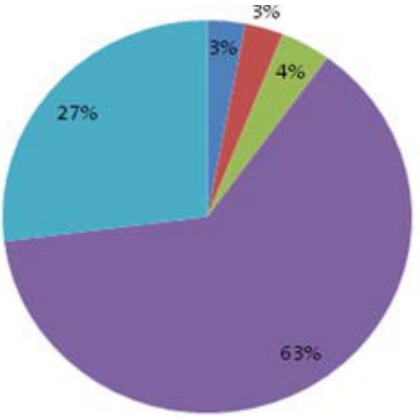

Fig. (4) category of customers

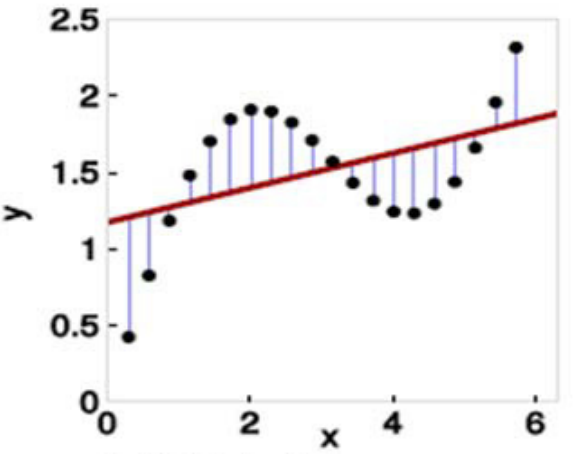

Fig (5) . Illustration of least squares

5.3. Errors Detection

Distance between observed values y_n and the predictions f(x_n), which are known as the residuals shown in (Fig5): black dots represents the training examples ,thick line is the function f(x) and vertical line represents residuals.

5.4 Results comparisons

In order to verify the performance of the prediction model mentioned in sub section 5.2, we use Eviews (SARIMA), SPSS (SARIMA) and Support Vector Machine (SVM) to predict different products' supply quantity. The relative error is chosen as the metrics, and is calculated as follows eq 8:

$$error\ rate = \frac{[F[m,j] - R[m,j]]}{R[m,j]} * 100\% \quad (8)$$

Where $F[m, j]$ depicts the predicted supply quantity of *j-th* month in the *m-th* year, $R[m, j]$ denotes the real supply quantity of *j-th* month in the *m-th* year as described in Table 2.

Table 2.error estimation rate and comparisons

Product name	SSPS	Eviews	SVM	Trend model
House 1	24%	53%	21%	13%
House 2	12%	41%	22%	17%

6.Conclusion

Marketing research has provided many interesting results that have not been adequately applied. This is due to the fact that the research has not considered the managers to be an important part of the study. When the estimates are drawn to support managers in their marketing-mix decisions, they cover not only the statistical issues of avoiding and fitting the bias but also yield the optimizable models. Hence the company can have good decision making capability.

Despite making a decision- making frame- work easy for managers, the model sometimes proves to be imperfect. The reasons behind this can be (i)the model can violate several requirements of a "proper" linear regression model. (ii)inadequate rows of data which prevent the making of the valid model. (iii) Its not always that every column in the model will be truly independent as in some cases negative coefficient can be seen in front of some variables, which do not make any sense at all.

Future scope for this model is not just about outputting a single number but also about identifying the patterns and rules. Model performs well while evaluating the randomly choose test data. It also provides a pretty good solution to many of the problems. This regression model is easy to use and can be used for big data sets. It gives the solution within no seconds. In case the model proves to be incorrect, one can still fix this by using preprocess tab and thus remove the particular column from the data set.

References

1. Chen, L. Y., & Wang, T. C. (2009), Optimizing partners' choice in IS/IT outsourcing projects:" The strategic decision of fuzzy" , *International Journal of Production Economics.*
2. Lin, C. T., Chen, C. B., & Ting, Y. C. (2011).,An ERP model for supplier selection in electronics industry. *Expert Systems with Applications*.
3. Hsu, B. M., Chiang, C. Y., & Shu, M. H. (2010), Supplier selection using fuzzy quality data and their applications to touch screen,. "*Expert Systems with Applications.*
4. Bronnenberg, B. J., Rossi, P. E., & Vilcassim, N. J. (2005)," Structural modeling and policy simulation," Journal of Marketing Research.
5. Franses, Ph. H. (2005)," Diagnostics, expectations, and endogeneity", Journal of Marketing Research.
6. Luan, Y. J., & Sudhir, K. (2010), "Forecasting marketing-mix responsiveness for new products", Journal of Marketing Research.
7. Tenopir/King, Towards Electronic Journals.
8. Joseph Hair, Rolph E. Anderson, Ronald L. Tatham, and William C. Black(1998)," Multivariate Data Analysis", (Upper Saddle River: Prentice-Hall).
9. John Cox and Laura Cox , (2006)., Scholarly Publishing Practice, " Policies and Practices in Online Publishing, Second Survey" Academic Journal Publishers (Brighton: Association of Learned and Professional Society Publishers).